BAYESIAN STATISTICS 7, pp. 25–43
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2003

Hierarchical Bayesian Models for Applications in Information Retrieval

DAVID M. BLEI, MICHAEL I. JORDAN and ANDREW Y. NG
University of California, Berkeley, USA
blei@cs.berkeley.edu jordan@cs.berkeley.edu ang@cs.berkeley.edu

SUMMARY

We present a simple hierarchical Bayesian approach to the modeling collections of texts and other large-scale data collections. For text collections, we posit that a document is generated by choosing a random set of multinomial probabilities for a set of possible "topics," and then repeatedly generating words by sampling from the topic mixture. This model is intractable for exact probabilistic inference, but approximate posterior probabilities and marginal likelihoods can be obtained via fast variational methods. We also present extensions to coupled models for joint text/image data and multiresolution models for topic hierarchies.

Keywords: VARIATIONAL INFERENCE METHODS; HIERARCHICAL BAYESIAN MODELS; EMPIRICAL BAYES; LATENT VARIABLE MODELS; INFORMATION RETRIEVAL.

1. INTRODUCTION

The field of information retrieval is broadly concerned with the problem of organizing collections of documents and other media so as to support various information requests on the part of users. The familiar problem of returning a subset of documents in response to a query lies within the scope of information retrieval, as do the problems of analyzing cross-referencing within a document collection, analyzing linguistic structure so as to be able to say who did what to whom, automatic compilation of document summaries, and automatic translation. Another class of problems involve analyzing interactions between users and a collection, including the "collaborative filtering" problem in which suggestions are made to new users of a system based on choices made by previous users.

Clearly there is much grist for the mill of Bayesian statistics in the information retrieval problem. One can view the information retrieval system as uncertain about the needs of the user, an uncertainty which can be reduced in an ongoing "learning" process via a dialog with the user or with a population of users. There are also many modeling issues—particularly those surrounding the appropriate level of resolution at which to view a document collection—where the tools of hierarchical Bayesian modeling are clearly appropriate. The problem of formulating a response to a query can be viewed in Bayesian terms: one can model the conditional distribution of queries given documents, and in conjunction with a model of documents treat the problem of responding to a query as an application of Bayes' theorem (Zhai and Lafferty, 2001). Finally, many problems in information retrieval involve preferences and choices, and decision-theoretic analysis is needed to manage the complexity of the tradeoffs that arise.

Despite these natural motivations, it is generally not the case that current information retrieval systems are built on the foundation of probabilistic modeling and inference. One important reason for this is the severe computational constraints of such systems. Collections involving tens or hundreds of thousands of documents are commonplace, and the methods that are used by information retrieval systems—reduction of a document to a "vector" of smoothed frequency counts, possibly followed by a singular value decomposition to reduce dimensionality, and computations of inner products between these "vectors"—have the important virtue of computational efficiency. If probabilistic modeling is to displace or augment such methods, it will have to be done without a major increase in computational load—and there are questions about whether this can be done with the current arsenal of Bayesian tools. A user who has sent a query to a search engine is generally not willing to wait for a Markov chain Monte Carlo simulation to converge.

In the current paper, we discuss a class of hierarchical Bayesian models for information retrieval. While simple, these models are rich enough as to yield intractable posterior distributions, and to maintain computational efficiency, we make use of *variational inference methods*. Variational methods yield deterministic approximations to likelihoods and posterior distributions that provide an alternative to Markov chain Monte Carlo. They are particularly apt in a domain such as information retrieval in which a fast approximate answer is generally more useful than a slow answer of greater fidelity.

The models that we discuss are all instances of so-called "bag-of-words" models (Baeza-Yates and Ribeiro-Neto, 1999). Viewing the words in a document as random variables, these are simply models in which the words are exchangeable. While clearly a drastic simplification, this assumption is generally deemed necessary in information retrieval because of the computational constraints; moreover, it has been found in practice that viable information retrieval systems can be built using such an assumption. In any case, by building up a sufficiently detailed model for the mixture underlying the word distribution, we hope to ameliorate the effect of the exchangeability assumption, and capture some of the latent structure of documents while maintaining computational tractability.

Finally, although we believe that probabilistic methods have an important role to play in information retrieval, full Bayesian computations are often precluded by the size of the problems in this domain. We therefore make significant use of empirical Bayesian techniques, in particular fixing hyperparameters via maximum likelihood. Some comfort is provided in this regard by the large scale of the problems that we study, but it also is important to acknowledge that all of the models that we study are very inaccurate reflections of the underlying linguistic reality.

2. LATENT VARIABLE MODELS OF TEXT

In any given document collection, we envision a number of underlying "topics"; for example, a collection may contain documents about novels, music or poetry. These topics may also viewed at varying levels of resolution; thus, we have have documents about romance novels, jazz music or Russian poetry. A traditional approach to treating such "topics" is via hierarchical clustering, in which one represents each document as a vector of word counts, defines a similarity measure on the vectors of word counts, and applies a hierarchical clustering procedure in the hopes of characterizing the "topics." A significant problem with such an approach, however, is the mutual exclusivity assumption—a given document can only belong to a single cluster. Textual material tends to resist mutual exclusivity—words can have several different (unrelated) meanings, and documents can be relevant to different topics (a document can be relevant to both jazz music and Russian poetry). If we are to use clustering procedures—and indeed computational concerns lead us to aim towards some form of divide-and-conquer strategy—care must be taken to define clustering procedures that are suitably flexible.

To provide the requisite flexibility, we propose a hierarchical Bayesian approach. The basic scheme is that of a mixture model, corresponding to an exchangeable distribution on words. The mixture has two basic levels. At the first level, we have a finite mixture whose mixture components can be viewed as representations of “topics.” At the second level, a latent Dirichlet variable provides a random set of mixing proportions for the underlying finite mixture. This Dirichlet variable can be viewed as a representation of “documents”; *i.e.*, a document is modeled as a collection of topic probabilities. The Dirichlet is sampled once per document, and the finite mixture is then sampled repeatedly, once for each word within a document. The components that are selected during this sampling process are a multiset that can be viewed as a “bag-of-topics” characterization of the document.

We describe this basic model in detail in the following section, and treat various extensions in the remainder of the paper.

3. LATENT DIRICHLET ALLOCATION

The basic entity in our model is the *word*, which we take to be a multinomial random variable ranging over the integers $\{1, \dots, V\}$, where V is the *vocabulary size*, a fixed constant. We represent this random variable as a V-vector w with components $w^i \in \{0, 1\}$, where one and only one component is equal to one.

A *document* is a sequence of N words, denoted by $\boldsymbol{w}$, where w_n is the n^{th} word. We assume that we are given as data a *corpus* of M documents: $\mathcal{D} = \{\boldsymbol{w}_d : d = 1, \dots, M\}$.

We refer to our model as a *Latent Dirichlet allocation (LDA)* model. The model assumes that each document in the corpus is generated as follows:

(1) Choose $N \sim p(N \mid \xi)$.
(2) Choose $\theta \sim \text{Di}(\alpha)$.
(3) For each of the N words w_n:
  (a) Choose a topic $z_n \sim \text{Mult}(\theta)$.
  (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionality k of the Dirichlet distribution (and thus the topic variable z) is assumed known and fixed. Second, the word probabilities are parameterized by a $k \times V$ matrix β where $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$, which for now we treat as a fixed quantity that is to be estimated. Finally, note that we have left the document length distribution unspecified—in most applications we work conditionally on N, and indeed in the remainder of the paper we generally omit reference to N.

Given this generative process, the joint distribution of a topic mixture θ, N topics $\boldsymbol{z}$, and an N word document $\boldsymbol{w}$ is:

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}, N \mid \alpha, \beta) = p(N \mid \xi)\, p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \beta), \tag{1}$$

where $p(z_n \mid \theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. This model is illustrated in Figure 1 (Left).

Note the distinction between LDA and a simple Dirichlet-multinomial clustering model. In the simple clustering model, the innermost plate would contain only w, the topic node would be sampled only once for each document, and the Dirichlet would be sampled only once for the whole collection. In LDA, the Dirichlet is sampled for each document, and the multinomial topic node is sampled *repeatedly* within the document.

4. INFERENCE

Let us consider the problem of computing the posterior distribution of the latent variables θ and z given a document (where we drop reference to the randomness in N for simplicity):

$$p(\theta, \boldsymbol{z} \mid \boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta)}{p(\boldsymbol{w} \mid \alpha, \beta)}.$$

We can find the denominator—a marginal likelihood—by marginalizing over the latent variables in Equation 1:

$$p(\boldsymbol{w} \mid \alpha, \beta) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j}\right) d\theta. \qquad (2)$$

This is an expectation under an extension to the Dirichlet distribution which can be represented with special hypergeometric functions (Dickey, 1983, Dickey, Jiang, and Kadane, 1987). Unfortunately, this function is infeasible to compute exactly, due to the coupling between θ and β inside the summation over latent factors.

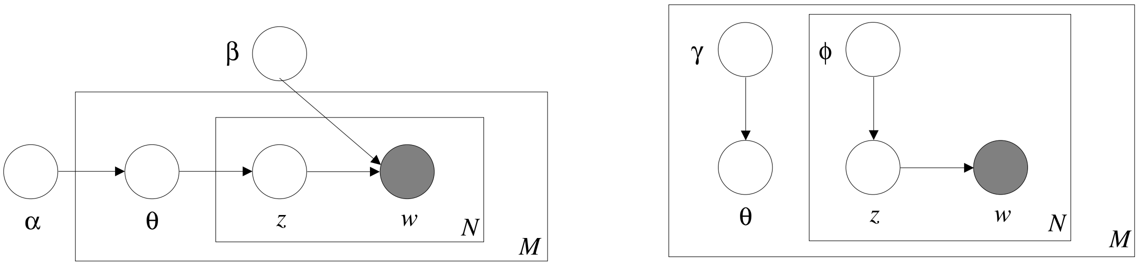

Figure 1. *(Left) Graphical model representation of LDA. The box is a "plate" representing replicates. (Right) Graphical model representation of the variational distribution to approximate the posterior in LDA.*

Note that we are treating the parameters α and β as fixed constants in Eq. (2); we show how to estimate their values using an empirical Bayes procedure in this section. In Section 6 we consider a fuller Bayesian model in which β is endowed with a Dirichlet distribution.

To approximate the posterior probability in a computationally efficient manner, we make use of *variational inference algorithms* (Jordan, *et al.,* 1999). Variational inference is related to importance sampling in its use of a simplified distribution to approximate the posterior. Rather than sampling from such a distribution, however, we consider a family of simplified distributions, parameterized by a set of *variational parameters*. Ranging over these parameters, we find the best approximation in the family, measuring approximation quality in terms of KL divergence. We essentially convert the inference problem (a problem of computing an integral) into an optimization problem (a problem of maximizing a function).

In the context of graphical models, the simplified distributions that provide the approximations used in a variational inference framework are generally obtained by omitting one or more edges in the graph. This decouples variables and provides a more tractable approximation to the posterior.

Figure 1 (Left) illustrates the LDA model for a corpus of documents. In this graph, the problematic coupling between θ and β is represented as the arc between θ and $\boldsymbol{z}$. We develop a variational approximation by defining an approximating family of distributions $q(\theta, \boldsymbol{z} \mid \boldsymbol{w}, \gamma, \phi)$, and choose the variational parameters γ and ϕ to yield a tight approximation to the true posterior.

In particular, we define the factorized variational distribution:

$$q(\theta, \boldsymbol{z} \,|\, \boldsymbol{w}, \gamma, \phi) = p(\theta \,|\, \boldsymbol{w}, \gamma) \prod_{n=1}^{N} p(z_n \,|\, \boldsymbol{w}, \phi_n),$$

as illustrated in Figure 1 (Right). This distribution has variational Dirichlet parameters γ and variational multinomial parameters ϕ_n. Note that all parameters are conditioned on $\boldsymbol{w}$; for each document, there is a different set of Dirichlet and multinomial variational parameters.

With this new model in hand, we can obtain an approximation to $p(\theta, \boldsymbol{z} \,|\, \boldsymbol{w}, \alpha, \beta)$ via a minimization of the KL divergence between the variational distribution and the true posterior:

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} D(q(\theta, \boldsymbol{z} \,|\, \boldsymbol{w}, \gamma, \phi) \parallel p(\theta, \boldsymbol{z} \,|\, \boldsymbol{w}, \alpha, \beta)). \tag{3}$$

As we show in the following section, we can take decreasing steps in the KL divergence and converge to (locally) optimizing parameters (γ^*, ϕ^*) by alternating between the following pair of update equations:

$$\phi_{ni} \propto \beta_{iw_n} \exp\{\mathrm{E}_q[\log(\theta_i \,|\, \gamma)]\} \tag{4}$$

$$\gamma_i = \alpha_i + \textstyle\sum_{n=1}^{N} \phi_{ni}, \tag{5}$$

where a closed-form expression for the expectation in Eq. (4) is given in Section 4.1 below. Note again that we obtain variational parameters (γ^*, ϕ^*) for each document $\boldsymbol{w}$.

Eqs. (4) and (5) have an appealing intuitive interpretation. The Dirichlet update in Eq. (5) yields a posterior Dirichlet given expected observations taken under the variational distribution. The multinomial update in Eq. (4) is akin to using Bayes' theorem, $p(z_n \,|\, w_n) \propto p(w_n \,|\, z_n)p(z_n)$, where $p(z_n)$ is approximated by the exponential of the expected value of its log under the variational distribution.

Each iteration of the algorithm requires $O(Nk)$ operations. Empirically, we find that the number of iterations required for a single document scales linearly in the number of words in the document. This yields a total number of operations that scales empirically as $O(N^2k)$.

4.1 *Variational inference*

In this section, we derive the variational inference equations in Eq. (4) and Eq. (5).

We begin by noting that the tranformed parameters, $\log \theta_i$, are the natural parameters in the exponential family representation of the Dirichlet distribution, and thus the expected value of $\log \theta_i$ is given by $\mathrm{E}[\log \theta_i \,|\, \alpha] = \Psi(\alpha_i) - \Psi\left(\sum_{i=1}^{k} \alpha_i\right)$, where $\Psi(x)$ is the digamma function.

To derive the variational inference algorithm, we begin by bounding the marginal likelihood of a document using Jensen's inequality (Jordan, *et al.*, 1999):

$$\begin{aligned}
\log p(\boldsymbol{w} \,|\, \alpha, \beta) &= \log \int_\theta \sum_{\boldsymbol{z}} p(\theta, \boldsymbol{z}, \boldsymbol{w} \,|\, \alpha, \beta)\, d\theta \\
&= \log \int_\theta \sum_{\boldsymbol{z}} \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w} \,|\, \alpha, \beta) q(\theta, \boldsymbol{z})}{q(\theta, \boldsymbol{z})} d\theta \\
&\geq \int_\theta \sum_{\boldsymbol{z}} q(\theta, \boldsymbol{z}) \log p(\theta, \boldsymbol{z}, \boldsymbol{w} \,|\, \alpha, \beta)\, d\theta - \int_\theta \sum_{\boldsymbol{z}} q(\theta, \boldsymbol{z}) \log q(\theta, \boldsymbol{z})\, d\theta \\
&= \mathrm{E}_q[\log p(\theta, \boldsymbol{z}, \boldsymbol{w} \,|\, \alpha, \beta)] - \mathrm{E}_q[\log q(\theta, \boldsymbol{z})].
\end{aligned} \tag{6}$$

It is straightforward to show that the difference between the left-hand side and the right-hand side of this equation is precisely the KL divergence in Eq. (3), and thus minimizing that KL divergence is equivalent to maximizing the lower bound on the marginal likelihood in Eq. (6).

Letting $\mathcal{L}(\gamma, \phi; \alpha, \beta)$ denote the right-hand side of Eq. (6), we have:

$$\begin{aligned}\mathcal{L}(\gamma, \phi; \alpha, \beta) = {} & \mathrm{E}_q[\log p(\theta \mid \alpha)] + \mathrm{E}_q[\log p(\boldsymbol{z} \mid \theta)] + \mathrm{E}_q[\log p(\boldsymbol{w} \mid \boldsymbol{z}, \beta)] \\ & - \mathrm{E}_q[\log q(\theta)] - \mathrm{E}_q[\log q(\boldsymbol{z})].\end{aligned} \tag{7}$$

Next, we write Eq. (7) in terms of the model parameters (α, β) and variational parameters (γ, ϕ). Each of the five lines expands one of the five terms:

$$\begin{aligned}\mathcal{L}(\gamma, \phi; \alpha, \beta) = {} & \log \Gamma\left(\textstyle\sum_{j=1}^{k} \alpha_j\right) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) \\ & + \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) \\ & + \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} w_n^j \phi_{ni} \log \beta_{ij} \\ & - \log \Gamma\left(\textstyle\sum_{j=1}^{k} \gamma_j\right) + \sum_{i=1}^{k} \log \Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) \\ & - \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}.\end{aligned} \tag{8}$$

In the following sections, we take derivatives with respect to this expression to obtain our variational inference algorithm.

Variational multinomial. We first maximize Eq. (8) with respect to ϕ_{ni}, the probability that the n^{th} word was generated by latent topic i. Observe that this is a constrained maximization since $\sum_{i=1}^{k} \phi_{ni} = 1$.

We form the Lagrangian by isolating the terms which contain ϕ_{ni} and adding the appropriate Lagrange multipliers. Let β_{iv} refer to $p(w_n^v = 1 \mid z^i = 1)$ for the appropriate v (and recall that each w_n is a V-vector with exactly one component equal to 1; we can select the unique v such that $w_n^v = 1$). We have:

$$\mathcal{L}_{[\phi_{ni}]} = \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) + \phi_{ni} \log \beta_{iv} - \phi_{ni} \log \phi_{ni} + \lambda_n\left(\textstyle\sum_{j=1}^{k} \phi_{nj} - 1\right).$$

Taking derivatives with respect to ϕ_{ni}, we obtain:

$$\frac{\partial \mathcal{L}}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right) + \log \beta_{iv} - \log \phi_{ni} - 1 + \lambda_n.$$

Setting this derivative to zero yields the maximized ϕ_{ni} (cf. Eq. (4)):

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right). \tag{9}$$

Variational Dirichlet. Next, we maximize Eq. (8) with respect to γ_i, the i^{th} component of the posterior Dirichlet parameter. The terms containing γ_i are:

$$\begin{aligned}\mathcal{L}_{[\gamma]} = &\sum_{i=1}^{k}(\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k}\gamma_j\right)\right) + \sum_{n=1}^{N}\phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k}\gamma_j\right)\right)\\ &- \log\Gamma\left(\textstyle\sum_{j=1}^{k}\gamma_j\right) + \log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k}\gamma_j\right)\right).\end{aligned}$$

This simplifies to:

$$\mathcal{L}_{[\gamma]} = \sum_{i=1}^{k}\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k}\gamma_j\right)\right)\left(\alpha_i + \textstyle\sum_{n=1}^{N}\phi_{ni} - \gamma_i\right) - \log\Gamma\left(\textstyle\sum_{j=1}^{k}\gamma_j\right) + \log\Gamma(\gamma_i).$$

We take the derivative with respect to γ_i:

$$\frac{\partial\mathcal{L}}{\partial\gamma_i} = \Psi'(\gamma_i)\left(\alpha_i + \textstyle\sum_{n=1}^{N}\phi_{ni} - \gamma_i\right) - \Psi'\left(\textstyle\sum_{j=1}^{k}\gamma_j\right)\displaystyle\sum_{j=1}^{k}\left(\alpha_j + \textstyle\sum_{n=1}^{N}\phi_{nj} - \gamma_j\right). \quad (10)$$

Setting this equation to zero yields a maximum at:

$$\gamma_i = \alpha_i + \textstyle\sum_{n=1}^{N}\phi_{ni}. \quad (11)$$

Since Eq. (11) depends on the variational multinomial ϕ, full variational inference requires alternating between Eqs. (9) and (11) until the bound on $p(\boldsymbol{w} \mid \alpha, \beta)$ converges.

4.2 *Estimation*

In this section we discuss approximate maximum likelihood estimation of the parameters α and β. Given a corpus of documents $\mathcal{D} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}$, we use a variational EM algorithm (EM with a variational E step) to find the parameters α and β which maximize a lower bound on the log marginal likelihood:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M}\log p(\boldsymbol{w}_d \mid \alpha, \beta).$$

As we have described above, the quantity $p(\boldsymbol{w} \mid \alpha, \beta)$ cannot be computed efficiently. However, we can bound the log likelihood using:

$$p(\boldsymbol{w}_d \mid \alpha, \beta) = \mathcal{L}(\gamma, \phi; \alpha, \beta) + D(q(\theta, \boldsymbol{z} \mid \boldsymbol{w}_d, \gamma, \phi) \parallel p(\theta, \boldsymbol{z} \mid \boldsymbol{w}_d, \alpha, \beta)), \quad (12)$$

which exhibits $\mathcal{L}(\gamma, \phi; \alpha, \beta)$ as a lower bound since the KL term is positive.

We now obtain a variational EM algorithm that repeats the following two steps until Eq. (12) converges:

(E) Find the setting of the variational parameters $\{\gamma_d, \phi_d : d \in \mathcal{D}\}$ which tighten the bound in Eq. (12) as much as possible. This is simply variational inference for each training document as described in the previous section.

(M) Maximize Eq. (12) with respect to the model parameters α and β. This corresponds to finding the maximum likelihood estimates with the approximate expected sufficient statistics computed in the E step.

To maximize with respect to β, we isolate terms and add Lagrange multipliers:

$$\mathcal{L}_{[\beta_{ij}]} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{dni} w_n^j \log \beta_{ij} + \sum_{i=1}^{k} \lambda_i \left(\sum_{j=1}^{V} \beta_{ij} - 1 \right).$$

We take the derivative with respect to β_{ij}, set it to zero, and find:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_n^j.$$

The terms which contain α are:

$$\mathcal{L}_{[\alpha]} = \sum_{d=1}^{M} \log \Gamma \left(\sum_{j=1}^{k} \alpha_j \right) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1) \left(\Psi(\gamma_{di}) - \Psi \left(\sum_{j=1}^{k} \gamma_{dj} \right) \right)$$

Taking the derivative with respect to α_i gives:

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = M \left(\Psi \left(\sum_{j=1}^{k} \alpha_j \right) - \Psi(\alpha_i) \right) + \sum_{d=1}^{M} \Psi(\gamma_{di}) - \Psi \left(\sum_{j=1}^{k} \gamma_{dj} \right).$$

Given the coupling between the derivatives for the different α_j, we use Newton-Raphson to find the maximal α. The Hessian has the following form:

$$\frac{\partial \mathcal{L}}{\partial \alpha_i \alpha_j} = M \left(\Psi' \left(\sum_{j=1}^{k} \alpha_j \right) - \delta(i,j) \Psi'(\alpha_i) \right)$$

and this form allows us to exploit the matrix inversion lemma to obtain a Newton-Raphson algorithm that requires only a linear number of operations (Ronning, 1989).

5. EXAMPLE

We illustrate how LDA works by examining the variational posterior parameters γ and ϕ_n for a document in the TREC AP corpus (Harman, 1992). Recall that ϕ_{nj} is an approximation to the posterior probability associated with the i^{th} topic and the n^{th} word. By examining $\max_i \phi_{ni}$, we obtain a proposed allocation of words to unobserved topics.

Furthermore, we can interpret the i^{th} Dirichlet parameter (maximized by Eq. (5)) as the i^{th} Dirichlet parameter for the model plus the expected number of instances of topic i which were seen in the given document. Therefore, subtracting the posterior Dirichlet parameters from the model Dirichlet parameters we obtain an indication of the degree to which each factor is present in a document.

We trained a 100-factor LDA model on a subset of the TREC AP corpus. The following is an article from the same collection on which we did not train:

> The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Figure 2. *The four factors with largest expected counts for an article from the AP corpus. We show the fifteen words with the largest probability, $p(w \mid z)$, for each of these factors.*

If we examine γ for this article, we find that most of the factors are very close to α while four of the factors achieve significant expected counts. Looking at the distribution over words, $p(w \mid z)$, for those four factors, we can identify the topics which mixed via the θ random variable to form this document (Figure 2).

6. SMOOTHING AND LDA

The large vocabulary size that is characteristic of many information retrieval problems creates serious problems of sparsity. A new document is very likely to contain words that did not appear in any of the documents in a training corpus. Maximum likelihood estimates of the multinomial parameters assign zero probability to such words, and thus zero probability to new documents. The standard approach to coping with this problem is to “smooth” the multinomial parameters, assigning positive probability to all vocabulary items whether or not they are observed in the training set (Jelinek, 1997). Laplace smoothing is commonly used; this corresponds to placing a uniform Dirichlet prior on the multinomial parameters.

Though it is often implemented in practice (e.g., Nigam, *et al.*, 1999), simple Laplace smoothing does not correspond to formal integration in the mixture model setting. In fact, exact integration is intractable for mixtures for the same reasons that exact inference is intractable under the LDA model. However, we can again utilize the variational framework and approximate the posterior Dirichlet given the data. We present this variational approximation in the remainder of this section.

We elaborate the basic LDA model to place Dirichlet priors on the parameters β_i, for $i \in \{1, \ldots, k\}$, where β_i are the multinomial probabilities $p(w \mid z^i = 1)$ of words given topics. Again making an exchangeability assumption, we have $\beta_i \sim \mathrm{Di}(\eta, \eta, \ldots, \eta)$. The probability of the data becomes:

$$p(\mathcal{D} \mid \alpha, \eta) = \int \prod_{i=1}^{k} p(\beta_i \mid \eta) \prod_{d=1}^{M} p(\boldsymbol{w}_d \mid \alpha, \beta)\, d\beta,$$

where $p(\boldsymbol{w} \mid \alpha, \beta)$ is simply the LDA model as described above.

This integral is intractable but we can again lower bound the log probability using a variational distribution:

$$\log p(\mathcal{D} \mid \alpha, \eta) \geq \mathrm{E}_q[\log p(\beta \mid \eta)] + \sum_{d=1}^{M} \mathrm{E}_q[\log p(\boldsymbol{w}_d, \boldsymbol{z}, \theta \mid \alpha, \beta)] + \mathrm{H(q)},$$

where the variational distribution takes the form:

$$q(\beta_{[1:K]}, \boldsymbol{z}_{[1:D]}, \theta_{[1:D]} \,|\, \boldsymbol{w}_{[1:D]}, \lambda, \phi, \gamma) = \prod_{i=1}^{K} \mathrm{Di}(\beta_i \,|\, \lambda_i) \prod_{d=1}^{M} q_d(\theta_d, \boldsymbol{z}_d \,|\, \boldsymbol{w}_d, \phi_d, \gamma_d).$$

Note that the variational parameter λ_i is a V vector (even though η is a scalar) and that q_d is the variational distribution for LDA as defined above.

Variational inference chooses values for the variational parameters so as to minimize the KL divergence between the variational posterior and the true posterior. The derivation is similar to the earlier derivation, and yields the following updates:

$$\begin{aligned} \lambda_{ij} &= \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} w_n^j \phi_{dni} \\ \gamma_{di} &= \alpha_i + \sum_{n=1}^{N_d} \phi_{dni} \qquad (13) \\ \phi_{dni} &\propto \exp\{\mathrm{E}[\log \beta_{iv} \,|\, \lambda_i]\} \exp\{\mathrm{E}[\log \theta_i \,|\, \gamma_d]\}, \qquad (14) \end{aligned}$$

where v in Eq. (14) refers to the unique index for which $w_{dn}^v = 1$.

To estimate values for the hyperparameters, we again maximize the lower bound on the log likelihood with respect to the expected sufficient statistics taken under the variational distribution. The maximization for α is the same as above. We calculate the following derivatives:

$$\begin{aligned} \frac{d \log p(\mathcal{D} \,|\, \alpha, \eta)}{d\eta} &= \sum_{i=1}^{K} \sum_{j=1}^{V} \mathrm{E}_q[\log \beta_{ij} \,|\, \lambda_i] + KV\Psi(V\eta) - KV\Psi(\eta) \\ \frac{d^2 \log p(\mathcal{D} \,|\, \alpha\eta)}{d\eta^2} &= KV^2\Psi'(V\eta) - KV\Psi'(\eta), \end{aligned}$$

and maximize η by Newton's method.

To compute the probability of a previously unseen document, we again form the variational lower bound:

$$\begin{aligned} \log p(\boldsymbol{w}_{\text{new}} \,|\, \alpha, \eta, \mathcal{D}) &\geq \mathrm{E}_q[p(\beta, \theta_{\text{new}}, \boldsymbol{z}_{\text{new}}, \boldsymbol{w}_{\text{new}} \,|\, \alpha, \eta, \mathcal{D})] \\ &= \mathrm{E}_q[p(\beta \,|\, \alpha, \eta, \mathcal{D})] + \mathrm{E}_q[p(\theta_{\text{new}}, \boldsymbol{z}_{\text{new}}, \boldsymbol{w}_{\text{new}})]. \end{aligned}$$

The optimizing parameters in the first term are exactly the λ_i computed in the empirical Bayes parameter estimation procedure. The optimizing parameters in the second term are found by simply iterating Eqs. (13) and (14) for the data in the new document.

7. EMPIRICAL RESULTS

In this section, we present an empirical evaluation of LDA on the benchmark CRAN corpus (van Rijsbergen and Croft, 1975), containing 2,630 medical abstracts with 7,747 unique terms, and a subset of the benchmark TREC AP corpus, containing 2,500 newswire articles with 37,871 unique terms. In both cases, we held out 10% of the data for test purposes and trained the models on the remaining 90%. Finally, note that in preprocessing all the data, we removed a standard list of stop words. (Further experimental details are provided in Blei, *et al.,* 2002).

We compared LDA to the following standard models: "unigram," "mixture of unigrams," and "pLSI." The "unigram" model is simply a single multinomial for all words, irrespective of the document. The "mixture of unigrams" is a finite mixture of multinomials (Nigam, *et al.,* 1999). The "probabilistic latent semantic indexing" (pLSI) model is a precursor of LDA in which the Dirichlet distribution of LDA is replaced with a list of multinomial probability vectors, one for each document in the corpus (Hofman, 1999). This is an over-parameterized model, and a "tempering" heuristic is used in practice to smooth the (maximum likelihood) solution. We fit all of the latent variable models using EM (variational EM for LDA) with exactly the same stopping criteria (the average change in expected log marginal likelihood is less than 0.001%).

To evaluate the predictive performance of these methods, we computed the *perplexity* of the held-out test set. The perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and can be thought of as the inverse of the per-word likelihood.

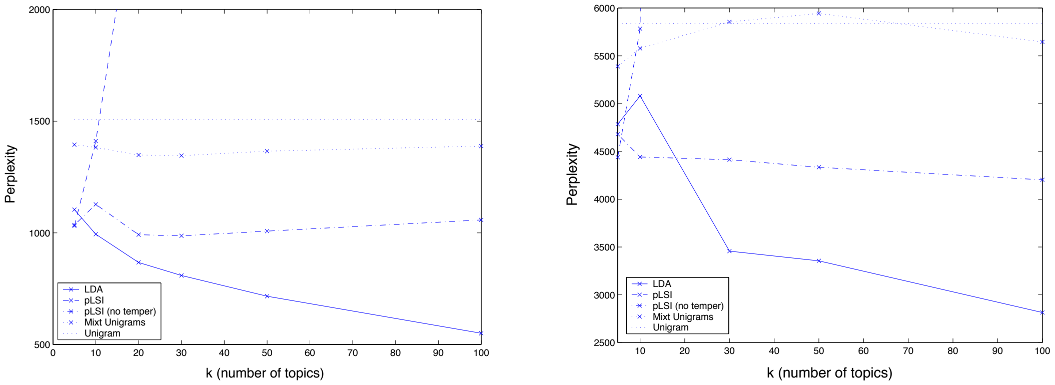

Figure 3. *Perplexity results on the CRAN (Left) and AP (Right) corpora for LDA, pLSI, mixture of unigrams, and the unigram model. Unigram is the higher dotted line; mixture of unigrams is the lower dotted line; untempered pLSI is dashed; tempered pLSI is dash-dot; and LDA is solid.*

Figure 3 illustrates the perplexity for each model and both corpora for different values of k. The latent variable models generally do better than the simple unigram model. The pLSI model severely overfits when not tempered (the values beyond $k = 10$ are off the graph) but manages to outperform the mixture of unigrams when tempered. LDA performs consistently better than the other models.

In Blei, *et al.* (2002) we present additional experiments comparing LDA to related mixture models in the domains of text classification and collaborative filtering.

8. EXTENSIONS

The LDA model is best viewed as a simple hierarchical module that can be elaborated to obtain more complex families of models for increasingly demanding tasks in information retrieval.

In applications to corpora involving sets of images that have been annotated with text, we have studied a model that we refer to as "Corr-LDA," consisting of two coupled LDA models (Blei and Jordan, 2002). As shown in Figure 4, the model has two kinds of observables—"words" and "image blobs"—and also has latent "topics" for both kinds of variables. Briefly, a Dirichlet variable is used to parameterize the mixing proportions for a latent topic variable for images. A correspondence between topic variables for images and words is enforced via an explicit "translational" conditional probability. This yields a model that can associate particular words with particular regions of the image. We show annotations in Figure 5, comparing Corr-LDA to "GM-Mixture," a joint mixture model for words and images, and "GM-LDA," a pair of

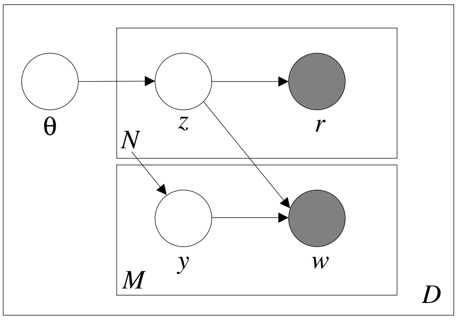

Figure 4. *The Corr-LDA model. The Gaussian random variable r encodes image blobs, while w encodes words. The variables z and y are latent "topics" for images and words, respectively. Note the "translational" conditional probability (the link from z to w) that enforces a correspondence between image topics and word topics.*

LDA models without the translational conditional. As suggested anecdotally by the annotations in the figure, and substantiated quantitatively in Blei and Jordan (2002), the Corr-LDA model is the most successful annotator of the models studied.

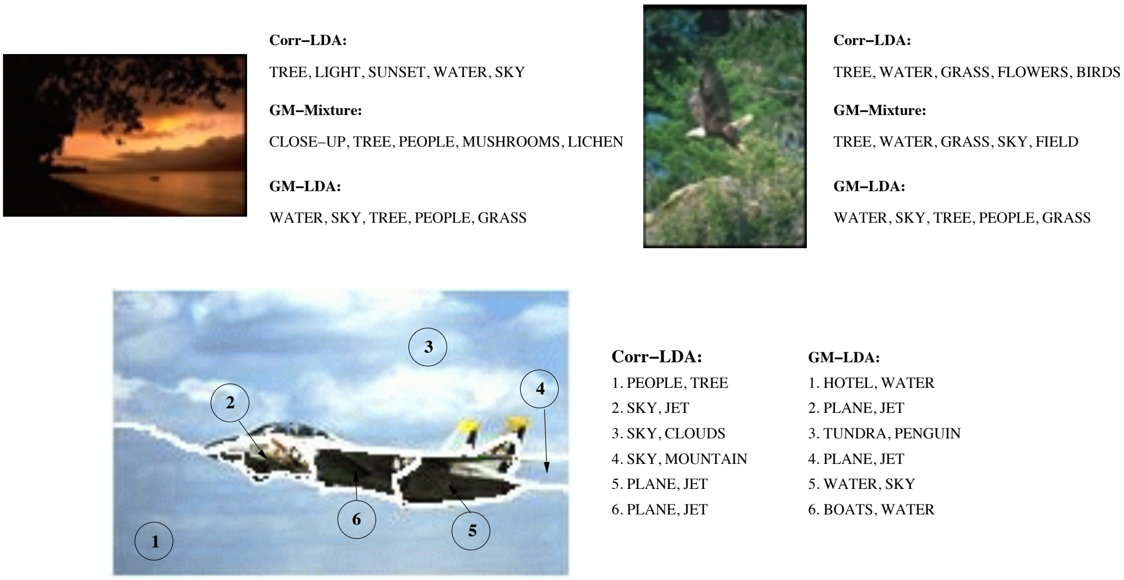

Figure 5. *(Top) Example images and their automatic annotations under different models. (Bottom) A segmented image and its labeling.*

Another important extension of LDA involves an attempt to capture a notion of "resolution." Thus, documents might discuss "music" in a broad sense, might specialize to talk about "classical music," and might further specialize to talk about "chamber music." As in our earlier discussion, we do not want to impose any mutual exclusivity assumption between such levels of resolution— we want to allow a single document to be able to utilize different levels at different times. Thus, some of the words in the document may be viewed as associated with very specific topics, while other words are viewed as more generic. Elaborating the LDA model, we obtain such a multiresolution model by consider a tree in which the nodes correspond to latent topics. The generative process for generating a document chooses a path in this tree, chooses a level along the path, and selects a topic from the corresponding nonterminal node. Under this process, documents will tend to share topics as they do in the basic LDA model, but will be particularly likely to share topics that are high in the topic hierarchy. Thus the hierarchy provides a flexible way to "share strength" between documents.

Both the pair LDA and the multiresolution LDA model are intractable for exact inference, but variational approximations are readily developed in both cases.

9. RELATED WORK

The LDA model and its hierarchical extension were inspired by the "probabilistic latent semantic indexing (pLSI)" model of Hoffman (1999). The pLSI model also assumes a mixture-of-multinomial model for the words in a document, but replaces the Dirichlet of LDA with a list of multinomial probability vectors, one for each document in the corpus. This can be viewed as a highly nonparametric version of LDA. As such, pLSI suffers from overfitting problems and is unable to assign probability mass to documents outside of the training corpus.

LDA is closely related to the Bayesian approach to mixture modeling developed by Diebolt and Robert (1994) and the Dirichlet multinomial allocation (DMA) model of Green and Richardson (2001). The DMA model posits a Dirichlet prior on a set of mixing proportions and a congugate prior on the parameters associated with each mixture component. The mixing proportions and component parameters are drawn once. Each data point is assumed to have been generated by first drawing a mixture component, and then drawing a value from the corresponding parameters.

The main differences between LDA and DMA arise from the role of the Dirichlet random variable as a representation of "documents" in the LDA setting. In LDA, we draw a set of mixing proportions multiple times from the Dirichlet; in DMA, the mixing proportions are drawn only once. For DMA, new data points are assumed to have been drawn from a single mixture component; in LDA, new data points are collections of words, and each word is allocated to an independently drawn mixture component.

10. DISCUSSION

We have presented a simple hierarchical approach to modeling text corpora and other large-scale collections. The model posits that a document is generated by choosing a random set of multinomial probabilities for a set of possible "topics," and then repeatedly generating words by sampling from the topic mixture. An important virtue of this model is that a given document can be characterized by several topics—we avoid the mutual exclusivity assumption that is made by most clustering models.

One limitation of our current work on LDA is that we assume that the number of topics is a user-defined parameter. While our empirical work has shown a lack of sensitivity to this parameter, it clearly is of interest to study inference methods (variational or MCMC; see, e.g., Attias, 2000, and Green and Richardson, 1998) that allow this parameter to be inferred from data.

While we have focused on applications to information retrieval, problems with modeling collections of sequences of discrete data arise in many other areas, notably bioinformatics. While clustering methods analogous to those used in information retrieval have been usefully employed in bioinformatics, the mutual exclusivity assumption underlying these methods is particularly unappealing in the biological setting, and it seems likely that LDA-style models based on "topics" can play a useful role in these problems.

ACKNOWLEDGEMENTS

We would like to acknowledge support from NSF grant IIS-9988642 and ONR MURI N00014-00-1-0637.

REFERENCES

Attias, H. (2000). A variational Bayesian framework for graphical models. *Advances in Neural Information Processing Systems* **12** (S. Solla, T. Leen and K-R. Mueller, eds.). Cambridge: MIT Press, 209–215.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

Blei, D. and Jordan, M. (2002). Modeling annotated data. *Tech. Rep.*, University of California, Berkeley, USA..

Blei, D., Jordan, M., and Ng. A. (2002). Latent Dirichlet allocation. *Journal of Machine Learning Research* (submitted).

Dickey, J. M. (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *J. Amer. Statist. Assoc.* **78**, 628–637.

Dickey, J. M., Jiang, J. M., and Kadane, J. B. (1987). Bayesian methods for censored categorical data. *J. Amer. Statist. Assoc.* **82**, 773–781.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. B* **56**, 363–375.

Green, P. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian J. Statist.* **28**, 355–377.

Harman, D. (1992). Overview of the first text retrieval conference (TREC-1). *Proceedings of the First Text Retrieval Conference* (D. Harman, ed.). NIST Special Publication, 1–20.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd. Annual International SIGIR Conference* (M. Hearst, F. Gey and R. Tong, eds.). New York: ACM Press, 50–57.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. Cambridge, MA: The MIT Press.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). Introduction to variational methods for graphical models. *Machine Learning* **37**, 183–233.

Nigam, K., Lafferty, J., and McCallum, A. (1999). Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering* (T. Joachims, A. McCallum, M. Sahami and L. Ungar, eds.). San Mateo: CA Morgan Kaufmann, 61–67.

Ronning, G. (1989). Maximum likelihood estimation of Dirichlet distributions. *J. Statist. Computation and Simulation* **34**, 215–221.

van Rijsbergen, C. and Croft, W. (1975). Document clustering: An evaluation of some experiments with the Cranfield 1400 collection. *Information Processing and Management* **11**, 71–182.

Zhai, C. and Lafferty, J. (2001). Document language models, query models, and risk minimization for information retrieval. *SIGIR Conference on Research and Development in Information Retrieval* (W. Croft, D. Harper, D. Kraft and J. Zobel, eds.). New York: ACM Press, 111–119.

DISCUSSION

STEVEN N. MACEACHERN *(The Ohio State University, USA)*

In this paper, Blei, Jordan and Ng have done a favor to the Bayesian statistical community by introducing the variational inference paradigm and by illustrating its benefits in modelling documents. The further illustration in the talk, of how well the technique works for the description of pictures in terms of a small set of words, was impressive. I suspect that this paradigm will become one of the standard methods for Bayesian analysis in problems where speed of computation is essential. My comments will focus on two issues–namely an overview of the authors' modelling strategy and a recommendation for what I call directional assessment and adjustment for Bayesian analysis. I look forward to the authors' views on whether a directional adjustment is appropriate in this context, whether such an adjustment would be generally appropriate for variational inference, and also whether such adjustment is compatible with variational inference.

The authors implement a three step strategy for their analysis: First, they choose simple, quickly computed statistics to monitor (each document is summarized by its word counts); second, they develop a simple model that generates these statistics (the multiple topics per document model, with conditionally independent choice of words drawn from the bag-of-words for each topic); third, they approximate the fit of the simple model (through the variational approximation). This strategy of writing a simplified model and then fitting it either exactly or by approximation is a mainstay of Bayesian inference. In many applications, one can provide a qualitative critique of a model. A simple model is retained for analysis, either because it simplifies computation or because specification of the structure and prior distribution for a more complex model would be difficult or not accepted by others. A great strength of this paper lies in its focus on computational algorithms that scale well and on the variational approximation used to fit the model. The next few paragraphs examine use of the strategy in a simplified context.

As an illustration of the contrast between simple and more complex models, consider two hierarchical models which have, at the middle stage, either a conditionally i.i.d. normal component or a stationary AR(1) component. The two models induce the same marginal distribution for each of the $\boldsymbol{Y}_i$. In the context of a larger model, with fairly large sample sizes, one would essentially learn the marginal distribution of the $\boldsymbol{Y}_i$, although their joint distribution could never be discovered from the simple analysis.

The simple model leads to inference for $\boldsymbol{\mu}$ based upon $\bar{\boldsymbol{Y}}$. This model has $\boldsymbol{\mu} \sim N(0, \boldsymbol{\sigma}_{\mu}{}^2)$; $\boldsymbol{\theta}_i \mid \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}_{\theta}{}^2)$; $\boldsymbol{Y}_i \mid \boldsymbol{\theta}_i \sim N(\boldsymbol{\theta}_i, \boldsymbol{\sigma}_{Y}{}^2)$, where $i = 1, \ldots, n$. The more complex model replaces the distribution on the $\boldsymbol{\theta}_i$ with a dependence structure, yielding $\boldsymbol{\theta}_0 \mid \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}_{\theta}{}^2)$; $\boldsymbol{\theta}_i \mid \boldsymbol{\mu}, \boldsymbol{\theta}_{<i} \sim N(\boldsymbol{\rho}\boldsymbol{\theta}_{i-1} + (1-\boldsymbol{\rho})\boldsymbol{\mu}, (1-\boldsymbol{\rho}^2)\boldsymbol{\sigma}_{\theta}{}^2)$, where $i = 1, \ldots, n$.

Since the sampling distribution of $\bar{\boldsymbol{Y}} \mid \boldsymbol{\mu}$ is normal under both models, the posterior distribution for $\boldsymbol{\mu} \mid \bar{\boldsymbol{Y}}$ can be found in closed form in both cases. In the simple model, the distribution of $\bar{\boldsymbol{Y}} \mid \boldsymbol{\mu}$ has mean $\boldsymbol{\mu}$ and variance $n^{-1}(\boldsymbol{\sigma}_{\theta}{}^2 + \boldsymbol{\sigma}_{Y}{}^2)$ while for the complex model it has mean $\boldsymbol{\mu}$ and variance $n^{-1}(\boldsymbol{\sigma}_{\theta}{}^2 + \boldsymbol{\sigma}_{Y}{}^2 + 2\boldsymbol{\sigma}_{\theta}{}^2\boldsymbol{\rho}/(1-\boldsymbol{\rho})) + O(n^{-2})$. The posterior distributions for $\boldsymbol{\mu} \mid \bar{\boldsymbol{Y}}$ under the models reflect this difference forever, with the ratio of posterior variances never tending to 1.

This overdispersion is not tied to the particulars of our simple and complex models. Instead, it is a general feature that appears whenever, loosely speaking, the $\boldsymbol{\theta}_i$ exhibit positive dependence. Thus, a qualitative assessment of the differences between the simple model and the complex model may lead us to conclude that the actual distribution of the statistic used in the Bayesian update has larger spread than the distribution we have used for formal calculation. As Bayesians, we are compelled to consider this assessment of our model and, if we judge the more complex posterior distribution to lie in a particular direction, to adjust our formally calculated posterior distribution.

There are a number of different perspectives that generate directional adjustments. A popular approach to adjustment is to flatten the likelihood either by raising it to a fractional power or by inflating the variance of a statistic. In the context of normal theory models, the approaches coincide. In the broader context of the exponential family, the adjustment can often be viewed as replacing the actual sample size with a smaller effective sample size while retaining the same values for the sufficient statistics. In early work on information processing, Zellner describes such adjustment, while Ibrahim and Chen (2000) describe it as a means of downweighting prior experiments.

There is a long tradition in survey sampling of directional adjustment. When a complex survey is administered, calculation of a standard error for an estimate can be difficult. The design based standard error accounts for features of the sampling design such as stratification and cluster sampling. Often, the very information needed for calculation of a standard error is unavailable to the analyst due to confidentiality constraints. In many surveys, standard errors are larger than those calculated by treating the data as a simple random sample. These difficulties have led to the notion of the "design effect", an inflator for the standard error. A design effect may be laboriously calculated for a number of estimators, taking into account the entire sample design. This design effect is then applied to inflate the standard errors of other estimators. The same notion of selecting an adjustment or of creating a distribution of adjustments can be ported over to Bayesian statistics through the device of a fractional likelihood. The analyst elicits, through whatever means, the fraction or the adjustment to posterior sample size.

Consider a directional assessment for the latent Dirichlet model, as applied to the information retrieval problem. The simple model is designed to work with the word counts for each of the documents. As a first step, we look at features of the problem which either suggest overdispersion or underdispersion of the word counts, relative to the behavior expected under the simple model.

To examine overdispersion/underdispersion in the context of the multinomial distribution, we first match the distributions on the cell means. Here, this is the number of words in the document times the cell probabilities. Overdispersion is naturally produced through a hierarchical structure where a vector of cell probabilities is drawn from some distribution and words are conditionally independent draws from these cell probabilities. Underdispersion is naturally obtained through stratification, where individual words are drawn as multinomials with differing vectors of cell probabilities. For stratification to reduce the dispersion, the different vectors of cell probabilities would be used in fixed proportion.

Turning to the documents, there are two main features that suggest overdispersion. The first is that language is an individual construct. Different individuals, even when writing on the same collection of topics, will consistently make different word choices. This effect is strong enough that it has been used to attribute authorship of documents to individuals. Mosteller and Wallace's (1984) examination of the Federalist papers relies on such a strategy.

The second feature which suggests overdispersion is the dynamic, ever-changing nature of language. New words enter the language (a check in the Merriam-Webster lists the word 'Bayesian' as only dating from 1961). Old words disappear, and usage frequency changes. Thus, word use provides information for dating documents. For collections of documents written over an extended period of time, changes in language lead to overdispersion. In the context of scientific documents, authors writing with different backgrounds will use a different vocabulary to describe many of the same concepts.

Certain features of the documents also suggest movement toward underdispersion. The main effects in this direction are tied to the structure of written language. The effects appear on the scale of entire documents, of paragraphs, and of sentences. At the level of the document, the classic training for writing an essay suggests that one introduce the topic, put in the guts of the essay, and then wrap things up with conclusions. Word choice is presumably different in these three parts of a document. This indicates the presence of stratification which leads in the direction of underdispersion. At a mid-level, a paragraph generally begins with a topic sentence, with details filling out the remaining sentences. Again, with differential word choice in the initial and remaining sentences of a paragraph, this suggests stratification. At the lowest level, English (and I believe most languages) tends to place the main information content of a sentence toward its beginning. Again, we have evidence of stratification. The grammatical structure of language, with a need for nouns, verbs and adjectives, also suggests stratification.

In my experience, the presence of both sorts of effects, those that move toward overdispersion and those that move toward underdispersion, is common. To complete a directional assessment, one must judge the relative sizes of the competing effects. My impression is that, in the document context, an individual's word choice will provide by far the strongest effect, leading to a net overdispersion. With this in mind, my belief would be that an exact analysis based on the simple model would produce likelihoods that are too sharp, leading to a posterior distribution that is too concentrated.

The next step in a directional assessment of the latent Dirichlet analysis is to judge the impact of the variational approximation to the posterior on the analysis. Here, my intuition is much weaker, as I have only worked through the approximation in some very simple cases. However, in the cases I have examined, the approximation systematically results in a distribution which is more concentrated than the distribution being approximated. The difference is tied to the asymmetry of the Kullback-Liebler divergence. The variational approximation reverses the roles of the usual "true" and "near" distribution that appear in asymptotic theory.

Taken together, the full model and the variational approximation to the simple model suggest that some adjustment to the posterior distribution is appropriate. Directional adjustment suggests flattening all, or part, of the approximate posterior. A systematic examination across problems might suggest how much flattening is appropriate.

Acknowledgement. This material is based upon work supported by the National Science Foundation under Award No. DMS-0072526.

REPLY TO THE DISCUSSION

We agree with Steven MacEachern that overdispersion is an important issue for the line of research that we have presented, and an important issue for the field of probabilistic information retrieval. There are at least three distinct reasons why the posterior that we obtain is likely to be overly concentrated with respect to the distributions of words that we attempt to model. First, our model is exceedingly simple, ignoring many linguistic phenomena that will tend to yield larger variability than our model accounts for, in particular the sequential phenomena referred to by MacEachern. Second, the empirical Bayes methodology that we use is known to yield underestimates of posterior variance. Third, the variational approach that we utilize for inference tends to yield approximating distributions that are overly concentrated. Let us briefly discuss each of these phenomena and outline possible solutions.

The exchangeability assumptions underlying LDA are aimed at computational simplicity, but clearly they are overly strong. A general goal of the field of information retrieval is to relax "bag-of-words" assumptions, and develop models that are closer to linguistic reality. We view hierarchical Bayesian methodology as providing a natural upgrade path. By introducing latent variables for linguistic concepts such as "sense" and "style," we can begin to capture sources of variability that are currently outside of our model. In particular, the latter concept might be naturally introduced as a discrete multinomial that conditions the Dirichlet in LDA—yielding a mixture of Dirichlet distributions in the topic simplex, and allowing us to move beyond the restrictive assumption that the corpus is captured by a single Dirichlet in the topic space. We can also introduce Markovian structure; for example, the topic variable for each word can be conditioned on the topic variable for the preceding word. Alternatively, we can consider Dirichlet/multinomial mixtures on subsequences of words ("n-grams") rather than single words. Finally, we can also make use of information available from parsing algorithms to provide conditioning variables for the LDA model, thereby capturing some of the stratification that, as discussed by MacEachern, is likely to lead to underdispersion. In all of these cases, however, while it is easy to specify natural extensions of LDA within the hierarchical Bayesian formalism, computational issues are of major concern. Any model that is aimed at applications in information retrieval must scale to tens or hundreds of thousands of documents, and inferential procedures must run in seconds.

Despite the large scale of problems in information retrieval, the data are also often sparse. Indeed, as we discussed in Section 6, document-by-word matrices tend to be sparse, and a new document is very likely to contain words that were not seen in the training corpus. This problem has been the subject of intense study, often inspired by theoretical work on species-sampling models and generally studied within a frequentist framework (Chen and Goodman, 1996). The variational smoothing method outlined in Section 6 is an approximate Bayesian solution to this problem.

While the discussion in Section 6 provided an example of the advantage of moving beyond an empirical Bayes inference method to a fuller hierarchical Bayesian approach (for the parameter β), we also find that computational considerations often weigh in favor of the simplicity offered by the empirical Bayes approach (cf. the parameter α in the LDA model). We expect that these considerations will be of increasing importance as we consider richer and more complex hierarchical Bayesian models, and thus we expect that empirical Bayes will continue to play an important role in this line of research. In this regard, it is important to note that the well-known fact that empirical Bayes leads to underestimates of the posterior variance (Carlin and Louis, 2000). In the context of overdispersion, this further reduction in variability may be particularly problematic. The corrections discussed in the empirical Bayes literature (*e.g.*, Kass and Steffey, 1989) may provide some relief.

Finally, the convexity-based variational techniques presented in Section 4 are known to be overly concentrated relative to the posterior distribution that they approximate. This difficulty has been addressed in a number of ways. One approach involves using higher-order approximations–Leisink and Kappen (2002) have presented a general methodology for converting low-order variational lower bounds into higher-order variational bounds. It is also possible to achieve higher accuracy by dispensing with the requirement of maintaining a bound, and indeed Minka and Lafferty (2002) have shown that improved inferential accuracy can be obtained for the LDA model via a higher-order variational technique known as "expectation propagation." Another general approach involves combining variational methods with sampling techniques to improve the accuracy of the variational distribution while maintaining its simple form (de Freitas, *et al.,* 2001; Ghahramani and Beal, 2000). For example, the variational distribution can serve as a proposal distribution for an importance sampler.

Table 1. *Perplexity of a held-out test set as a function of the scaling factor.*

Scaling	Perplexity
0.1	2046.0
0.2	1818.6
0.3	1741.7
0.4	1661.1
0.5	1649.2
0.6	1652.1
0.7	1654.3
0.8	1654.0
0.9	1658.5
1.0	1660.5

While we believe that research in all of these areas—extended hierarchical Bayesian modeling of documents, corrections to empirical Bayes, and more accurate variational approximations—will eventually lead to well-motivated, computationally-efficient approximation procedures, the complexity of calibrating these various contributions to inaccurate assessment of variability also suggests that the kinds of "directional adjustments" suggested by MacEachern will play an important role. In an initial experiment, we adopted MacEachern's suggestion and rescaled the sample size while keeping the sufficient statistics fixed. Using a corpus of 740 documents, we estimated a five-factor LDA model, starting each run of variational EM from the same parameter setting. The results in Table 1 show the perplexity of a held-out set of 240 documents as a function of the scaling factor. The unscaled model was poorer than some of the scaled models, with the scaling of 0.5 yielding the best performance. Although these are preliminary results on a small corpus, they do indicate that simple directional adjustments may be useful in the Bayesian modeling of text corpora.

ADDITIONAL REFERENCES IN THE DISCUSSION

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman and Hall.

Chen, S. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics* (A. Joshi and M. Palmer, eds.). San Mateo, CA: Morgan Kaufmann, 310–318.

de Freitas, N., Højen-Sørensen, P., Jordan, M. I., and Russell, S. (2001). Variational MCMC. *Uncertainty in Artificial Intelligence* (J. Breese and D. Koller, eds.). San Mateo, CA: Morgan Kaufmann, 120–127.

Ghahramani, Z. and Beal, M. (2000). Variational inference for Bayesian mixtures of factor analyzers. *Advances in Neural Information Processing* **12**, Cambridge, MA: The MIT Press, 449–455.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15**, 46–60.

Kass, R. E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.* **90**, 773–795.

Leisink, M. and Kappen, H. (2002). General lower bounds based on computer generated higher order expansions. *Uncertainty in Artificial Intelligence*, (A. Darwiche and N. Friedman, eds.). San Mateo, CA: Morgan Kaufmann, 293–300.

Minka, T. and Lafferty, J. (2002). Expectation-propagation for the generative aspect model, *Uncertainty in Artificial Intelligence*, (A. Darwiche and N. Friedman, eds.). San Mateo, CA: Morgan Kaufmann, 352–359.

Mosteller, F. and Wallace, D. L. (1984). *Applied Bayesian and Classical Inference: The Case of* The Federalist *Papers*. New York: Springer